iS
CHANNEL

How Big Data Impacts Research and Knowledge Generation: An Epistemological View

Yiduo Wang

MSc in Information Systems and Digital Innovation
Department of Management
London School of Economics and Political Science

KEYWORDS

Literature Review
Big Data
Epistemology
Knowledge Generation
Prediction
Research
Social Sciences

ABSTRACT

Big data is a prevalent topic in today's news and articles, along with the opportunities it has created for academic research and business activities. However, it is difficult to analyse the real impact without an understanding of how big data has changed the way we generate knowledge. Therefore this review adopts an epistemological view. It critically engages with the literature and summarises two dominating themes. First, big data enables a new way of proposing theories by recognizing patterns purely from data. Second, the expansiveness of big data empowers large scale predictions in many areas. This review juxtaposes assumptions both for and against each argument and concludes that while big data has indeed created new disciplines and research paradigms, it is not a panacea for all the problems. Rather, asking the right questions and employing the appropriate methods are still critical to scientific discovery and value creation. Regarding literature selection, the author started from several Information Systems and Media and Communication journal articles and looked at their references for relevant literature, particularly those with keywords "big data", "epistemology", "theory" and "prediction". A group of articles are critically selected to give a comprehensive overview of the topic.

Introduction

The rise of big data has been much heralded in recent years, in company with technological progress such as increased computing power and the Internet of Things (IoT). Big data is not simply a technological advancement but has revolutionized business activities, academic research, human relationships and social interactions - namely every aspect of society. This critical literature review focuses on the epistemological discussions of big data. Epistemology, which originates from the Greek word epistēmē meaning "knowledge", is defined as "the study or a theory of the nature and grounds of knowledge especially with reference to its limits and validity" (Merriam-Webster's dictionary, n.d.). Therefore, the following review highlights how big data has changed the way academics and business generate knowledge as well as the validity of methodology used for big data analysis.

This article is structured as follows: first, big data is defined, with reasons why it differs from "small" data and accounts for a new epistemological change. Then, two categorizations from the literature are identified and introduced: one is the changing format of theory generation, the other is the unprecedented predictive power big data has created. Within both themes, views from proponents and detractors are carefully reviewed and validated with examples from various disciplines. Finally, this review ends with concluding remarks and possible areas for improvement and future research.

Defining Big Data: A Historical Review on Epistemology Changes

There are different views regarding how big data is defined. One school looks at big data comprehensively, identifying the scope and boundaries of big data from different perspectives. Ekbia et al. (2015) identified four main perspectives in the existing literature. The first perspective is product-oriented with a focus on the attributes of data, such as massiveness in volume, and data format such as audio or video. This perspective is adopted by Baesens et al. (2016) who defined the five "V" s of big data: volume, velocity, variety, veracity and value. The second perspective is process-oriented and underscores the novelty of processes that are required to analyse big data. For instance, the requisite technological infrastructure, tools and programming techniques advance with the emergence of both structured and unstructured data in various forms, such as text, audio, video and clickstream. The third perspective is cognition-oriented and highlights the concern that the human mind is no longer able to make sense of the large amount of data. This includes "capacity to search, aggregate, and cross-reference large data sets" (Boyd and Crawford, 2012, p.665). Finally, the fourth perspective, which was not explicitly mentioned in previous literature,

Corresponding Author
Email Address: yiduowang1234@outlook.com

is the social movement perspective. The emphasis is the socio-technical impact technology (big data) has on wider society, including economics, politics and culture. This view sees big data technologies as developed within a complex ecosystem formed by technology companies, the open source community, governments, and universities. For example, Yahoo! Supported the development of Apache Hadoop, the widely adopted open source framework in big data research, and IBM collaborated with universities to set up Data Science and Business Analytics programmes (Ekbia et al., 2015).

Other authors focus on the more distinctive feature of big data and argued that largeness in terms of size is not the main development (Chandler, 2015; Mills, 2018). Historically, there have been datasets larger than those currently regarded as big data, such as census data (Boyd and Crawford, 2012) and diary studies. One example, reported by Mills (2018), is the International Time-Use Study of 1965 by Szalai (1972), with 2,000 interviewees, aged 18-64 from 12 countries. Big data only renders the manual recording method obsolete but has no significant change in terms of data volume or time span. Rather, the distinction is that big data is not collected by researchers or governments to test a theory or validate a hypothesis, but is automatically generated from social media, mobile technologies, IoT, and on the internet. Therefore, data analysts seek to gain insights from data that already exists (Chandler, 2015). Such a view introduces the first theme of this literature review.

Regarding the changes big data creates, two themes dominate. One is that the vast amount of data provides an agnostic and comprehensive source of evidence, and therefore may change the way theories are proposed, tested and validated. This shift calls into question the correlation and causation between variables, which introduces a second topic regarding the predictive power of big data. That is, data reveals insights and predicts future trends even when the underlying mechanism is not clearly understood. Of course, both suffer from flaws, and these will be discussed in greater detail in the rest of this paper.

New Forms of Inquiry: "Data Speaks for Itself"

The vast amount of complex and relational datasets coupled with data analytics' techniques have challenged epistemologies in disciplines across the sciences, social sciences and humanities (Kitchin, 2014). Instead of collecting appropriate data for the sake of validating hypotheses and theories, researchers use data generated automatically from everyday behaviour. With the same input datasets, decisions such as which variables to count, which data to clean and what algorithms and models to employ lead to different results which sometimes engender unexpected discoveries (Dhar, 2013; Ekbia et al., 2015). Consequently, the computer is no longer "a pure analytic servant" but "an active question asking machine" (Agarwal and Dhar, 2014, p.444). Kitchin (2014) described this as a "new forms of empiricism" (p.1); that is, an epistemological approach for making sense of the world that is enabled by big data analysis. Rather than testing a theory by gathering relevant data, insights are acquired "born from the data" (ibid., p.2).

This shift in research paradigms is seen as a huge opportunity, or even a complete epistemic change towards an empiricism in which knowledge and patterns emerge from data themselves. One provocative forecast is voiced by Anderson (2008), stating that the scientific method of "hypothesize, model, test" is obsolete due to the deluge of data. He validated his argument by using the example of Craig Venter, who discovered thousands of previously unknown species of bacteria and other life-forms by statistically analysing and comparing large amounts of gene sequence data detected in the ocean and air, without knowing much of the new species. Such views are criticized fiercely by Pigliucci (2009), a philosopher of science, claiming that Anderson (2008) does not understand science and scientific methods. Although finding patterns is part of the scientific method, science is more about explanations for those patterns. Therefore, Venter's finding is just a starting point to form hypotheses. Without hypotheses to be tested, the data are just a "useless curiosity" (Pigliucci 2009, p.534). Even in a business scenario, advertisers are interested in theories of human behaviour and those theories act as guidance when making decisions about which data are collected and which keywords are used to organise the search. Moreover, Kitchin (2014) and Lazer et al. (2014) argue that the ability to recognise patterns also stems from previous scientific discoveries when theories are tested for validity and veracity. Thus, big data does not come out of a scientific vacuum but are part of a cumulative endeavour.

Besides the above debate regarding whether data is generated free from theory, the efficacy of such an inductive and empiricist scientific discovery approach pre-assumes some ideas underpinning its formulation, which could be fallacious (Kitchin, 2014). Two assumptions are summarized in the literature.

First, big data seeks to be exhaustive so that full resolution of the worldwide affairs can be captured (Steadman, 2013). However, data represents only parts of the population. For example, Boyd and Crawford (2012) pointed out that "people" and "Twitter users" are not synonymous and Twitter does not represent "all people" (p.669). Similarly, Floridi (2012) argued that the real epistemological problem with big data is the "small patterns" generated from pieces of data (p.436). Given that so much data can now be generated and processed so quickly and cheaply and on virtually anything, the pressure is to identify real value-adding patterns from the immense database. Such patterns, if found, only represent parts of the truth and would only be significant if aggregated properly. The requirement of aggregation and sense-making introduces the difficulty of integrating multiple data sources, both due to the constraints of computational power and the need to spot what has value in the data noise.

This leads to the second assumption that with big data, context or domain-specific knowledge is no longer needed, or is needed very little, in order to interpret

the data statistically (Anderson, 2008; Steadman, 2013). In order to recognize value from data noise, Floridi (2012) argues, techniques and technologies do help but are insufficient. Some data and computer scientists are active in practicing social science research are prone to "big data hubris" (Lazer et al., 2014, p.1203). Kitchin (2014) cites an example when a group of physicists employed big data analytics to model social and spatial processes in cities, hence suggested laws underpinning the process of city formation. He was critical that such studies often ignore both century-long social science practice and the effect of culture, politics and capital. From this point of view, it seems that the epistemological impact of big data is not fundamentally different to other new technologies which have changed measurement in scientific research (Kitchin 2014). The persistent problems remain; as Floridi (2012) quoted Plato (Cratylus, 390c), the crucial problem is "know how to ask and answer questions".

Predict from Big data: Casual Relations versus Statistical Correlations

Data-intensive disciplines and corresponding techniques can be traced backed to the 18th century with the development of statistics. There have long been debates about the difference between correlation and causal relationships, albeit less intensively, between the proponents of data-driven science and those of theory-driven science (Hey et al, 2009, cited by Ekbia et al. 2015). The same discourse continues to the big data age, when the deluge of economic and social transactions online make data much easier to access and make it easier to discover correlations among variables. The strain between correlational analysis and causal testing of hypotheses introduces the differences in the explanatory versus predictive power of big data. As has been argued by the philosopher of science Karl Popper, "prediction is a key epistemic criterion for assessing how seriously we should entertain a theory of a new insight: a good theory makes bold predictions that stand repeated effects as falsification." (Popper 1963, cited by Agarwal and Dhar (2014). Therefore, predictive power could be one of the strengths of big data.

However, some examples in the literature revealed different stories. Prediction using large-scale online data faces inherent difficulties when it goes beyond describing phenomena and tries to generalize public behaviour. For instance, Ekbia et al. (2015) reported the Emotive project where British researchers used Twitter and other social media data to map the emotions of the nation. Two thousand tweets were analysed per second and each tweet was categorised into eight human emotions (anger, disgust, fear, happiness, sadness, surprise, shame, and confusion). The researchers claimed their results could "help calm civil unrest and identify early threats to public safety" (BBC, 2013, para. 3). Nevertheless, Ekbia et al. (2015) questioned the validity of this prediction in two aspects. Socially, it is unclear to what extent this 'threat identification' is valuable for law reinforcement and under what context this will lead to order rather than chaos. Technically, two assumptions restrict the veracity of the result. First, human emotions can be reduced meaningfully to only eight categories, ignoring more subtle ones such as grief and contentment. Second, within the same category, emotions are expressed broadly without a distinction between, for example, happiness in different situations. Accordingly, these limitations may erode the credibility of the proposed prediction system.

More examples seem to support this view and raise concern over prediction using big data. Lazer et al. (2014) used the Google Flu Trends (GFT) example when Google tried to predict the number of doctor visits for influenza-like illnesses by key-word searching from 2009 to 2014. Even with improved models, the prediction is still two times higher than the actual record from Centers for Disease Control and Prevention (CDC) (Lazer et al., 2014). They analysed the causes for this deviated prediction and concluded with two reasons, namely "big data hubris" (Lazer et al., 2014, p.1203) and "algorithm dynamics (ibid., p.1203)". The assumption of the former is that big data are not a supplement to, but substitute for, traditional data aggregation and evaluation. This is a common hypothesis seen in big data related literature (Chandler, 2015; Dhar, 2013) and originates from the conceit of data and computer scientists who may practice social science without certain domain knowledge (Ekbia et al., 2015), as discussed before.

Therefore, the author calls for an "all data revolution" rather than a "big data revolution" (Lazer et al., 2014, p.1203) which emphasizes a combination of the traditional statistical methods with the new big data methodology. "Algorithm dynamics" (ibid., p.1203) mean that the algorithms alter in accordance with the business model of the commercial companies. That is, during the time span of the GFT project, Google also changed the data generating process (its algorithm) to improve customer service. While GFT takes in the assumption that the search frequency for certain terms is related to, and can reflect, external events, search behaviour was co-determined by exogenous determinants (such as user behaviour) as well as endogenous mechanisms (such as different algorithm models). "Algorithm dynamics" (ibid., p.1203) are also seen in other platforms such as Twitter and Facebook; since service providers kept re-engineering the algorithm, it was almost impossible to replicate the results. This poses an open question on the duplication ability of such big data research (Lazer et al., 2014) and contradicts with Kitchin (2014)'s proposition that the extensiveness of data makes it easier to test the veracity of theories.

The above two examples show what happens when the purpose of data-generating and data-analysing organizations (such as Twitter, Google and Facebook who emphasise profit and revenue) does not align with the purpose of what they are predicting (usually public affairs with the intention to enhance social welfare). Nevertheless, in business areas, big data prediction is valuable in terms of revenue generation and customer retention. Baesens et al. (2016) introduced several business use cases in their article, and one example was to use behaviour data to improve targeted marketing. Fine-grained transaction

data were analysed to predict which financial products individual customers were most likely to buy, and the results were not only of great value but also had higher quality with bigger data. The above instances highlight the urgency of selecting big data methods in the appropriate context. In short, big data is a powerful tool and can indeed provide insights in some circumstances, but the predictive power of big data again depends highly on what questions are asked and what context is studied.

Concluding Remarks

To conclude, this article recognized two distinctive features of big data, namely a new way of generating theories and the power to predict numerically from models and algorithms. Thus, it is pivotal to understand the question being asked so that appropriate data sets can be used to draw conclusions and insights.

This critical literature review could be further improved in terms of depth and breadth. It focused on a contemporary topic, in which most referenced articles are concentrated in the last decade. As there have been several shifts in research paradigms, it is worth putting big data into a historical context and comparing it with other scientific milestones such as the big science era after World War II. Also, for each of the two topics discussed, there are books worth looking at which would add more perspectives to the debate. For instance, Raw Data Is an Oxymoron questions the claims made about the objectivity of big data, and The Signal and the Noise: Why So Many Predictions Fail – but Some Don't would serve as good complementary reading for the prediction discourse. For breath, a third topic regarding big data and qualitative research is mentioned repeatedly in the literature but not elaborated on in the main text of this critical literature review due to time and word limits. Briefly, the massive amount of easily accessible data has also captured the imagination of qualitative researchers and several pieces of literature have mentioned the threats to, and potential of, big data in qualitative research (Ekbia et al. 2015, Mills 2018, Parks 2014). Interdisciplinary studies, such as computational social science and digital humanities, are also being developed due to big data, and would generate interesting discussion.

References

Anderson, C., 2008. The end of theory: The data deluge makes the scientific method obsolete. Wired, [online] 23 June 2008. Available at: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed 06 December 2019).

Agarwal, R., & Dhar, V. (2014). Editorial —Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. Information Systems Research, 25, 443–448. https://doi.org/10.1287/isre.2014.0546

Baesens, B., Bapna, R., Marsden, J. R., Vanthienen, J., & Zhao, J. L. (2016). Transformational Issues of Big Data and Analytics in Networked Business. MIS Quarterly., 40(4), 807–818. https://doi.org/10.25300/MISQ/2016/40:4.03

BBC, 2013. Computer program uses Twitter to 'map mood of nation'. BBC, [online] 7 September 2013. Available at: http://www.bbc.co.uk/news/technology-24001692 (accessed 13 December 2019).

Boyd, Danah, & Crawford, K. (2012). Critical Questions for Big Data. Information, Communication & Society, 15(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Chandler, D. (2015). A World without Causation: Big Data and the Coming of Age of Posthumanism, Millennium, 43(3), 833-851. https://doi.org/10.1177/0305829815576817

Dhar, V. (2013). Data Science and Prediction. Communication ACM, 56(12), 64–73. https://doi.org/10.1145/2500499

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazi, A., Bowman, T., Sugimoto, C. (2015). Big Data, Bigger Dilemmas: A Critical Review. Journal of the Association for Information Science and Technology, 66(8). 1523-1545. https://doi.org/10.1002/asi.23294

Epistemology. (n.d.) In Merriam-Webster's dictionary. Retrieved from https://www.merriam-webster.com/dictionary/epistemology

Floridi, L. (2012). Big Data and Their Epistemological Challenge. Philosophy & Technology, 25(4), 435–437. https://doi.org/10.1007/s13347-012-0093-4

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 2053951714528481. https://doi.org/10.1177/2053951714528481

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. Science, 343(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Mills, K. A. (2018). What are the threats and potentials of big data for qualitative research? Qualitative Research, 18(6), 591–603. https://doi.org/10.1177/1468794117743465

Parks, M. R. (2014). Big Data in Communication Research: Its Contents and Discontents. Journal of Communication, 64(2), 355–360. https://doi.org/10.1111/jcom.12090

Pigliucci, M. (2009). The end of theory in science? EMBO Reports, 10(6), 534. https://doi.org/10.1038/embor.2009.111

Steadman, I., 2013. Big data and the death of the theorist. Wired, [online] 25 January 2013. Available at: http://www.wired.co.uk/news/archive/2013-01/25/big-data-end-of-theory (accessed 10 December 2019).